ALIGNMENT-FREE GEOMETRY OF REPRESENTATIONAL TRAJECTORIES ACROSS THE BRAIN AND FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce an alignment-free Riemannian framework for comparing representational trajectories across systems that share neither coordinates, dimensionality, nor a common axis of evolution. We validate it by analyzing the response of the visual cortex and foundation models on common images. We study EEG and MEG responses to natural images from THINGS-EEG2 and THINGS-MEG, together with a label-fed language model, Llama-3-8B, and an image-fed vision model, DINOv2. For each system, an image is represented by a trajectory of neighborhood covariances on the symmetric positive-definite manifold. Each system's covariance trajectories are mapped to the tangent space at that system's Fréchet mean, velocity directions are computed along the system-specific axis, and systems are compared only through the pairwise angular geometry of per-image velocity directions. This statistic is invariant to an independent invertible congruence of each system's PCA-reduced covariances and to per-image rescaling. Across both EEG and MEG, cortical trajectories share positive angular geometry with both Llama and DINOv2. Order-sensitive tests separate two kinds of correspondence: the best-matching depth does not advance steadily in either model, while a rate-free monotone test finds cortical time and model depth ordered against the label-fed model in both cohorts and not against the image-fed one. Category analysis shows the correspondence is structured by graded neighborhood relations rather than reducible to discrete category labels, while scaffold controls reveal a substantial fraction of slice-wise concordance is induced by the shared stimulus-neighborhood graph. Methodologically, the framework separates coordinate-free geometric correspondence from order-preserving dynamical correspondence, providing a general tool to test when two representational systems share structure and trajectory.

1 INTRODUCTION

Given two systems exposed to the same inputs, a fundamental question is whether they organize and transform those inputs in similar ways. Direct comparison is not possible, since each system operates in its own representational space, with potentially different coordinates, dimensionality, and axis of evolution. A meaningful comparison should therefore be coordinate-free and relational, characterizing how each system organizes stimuli relative to one another rather than aligning individual representation dimensions. Moreover, because these representations evolve along system-specific axes, the comparison must be dynamical, capturing not only the representational structure at individual points but also how that structure changes along the respective trajectories. As a specific instance of our study, we examine the cortex (EEG/MEG) and two foundation models (image/text). Previous work has examined such correspondence one slice at a time: how similar are two representations at brain time t and model layer ℓ? We go beyond such fixed representations through explicit modeling of the dynamics. We represent each system's response to an image as a trajectory of neighborhood covariance matrices on the SPD manifold under the affine-invariant Riemannian metric, consider tangent spaces and the Fréchet means, and compare systems only through the relationship between the stimuli velocity directions. This relationship is provably unchanged by an independent invertible congruence of either system's covariances and by per-image rescaling. We

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Main findings across two recording modalities and two foundation models. Values are group means over 10 EEG or 4 MEG subjects, using a single neighborhood covariance over the occipito-temporal sensor array. Angle matrices are standardized separately within each subject and model (Eq. 6) and compared using Eq. 7. Bracketed fractions count subjects with the indicated sign: "pos." denotes a positive angular correlation or depth ridge, and "neg." a warped-path score with $z < 0$ relative to the depth-permutation null. Row 1 shows shared angular geometry in every subject, although peak depth does not replicate across cohorts: DINOv2 peaks at mid-depth in EEG and at 65% in MEG, Llama at 94% in EEG and at 35% in MEG. Row 2 shows weak order-preserving correspondence: depth ridges are weak and none is significantly positive, while the rate-free monotone test favors Llama in both cohorts. Rows 3–4 characterize the category structure and bound the contribution of the shared stimulus-neighborhood graph. Row 4 reports held-out first canonical correlations above and rank (Spearman) concordances of velocity directions below; these quantities are on different scales.

Finding / measure	EEG (n=10)	MEG (n=4)
1. Per-slice angular geometry is shared		
Llama-3-8B, label-fed	$+0.2632$ [10/10 pos.]	$+0.2835$ [4/4 pos.]
DINOv2 ViT-L/14, image-fed	$+0.3913$ [10/10 pos.]	$+0.3189$ [4/4 pos.]
Peak depth: DINOv2 / Llama	48% / 94%	65% / 35%
2. Order-preserving correspondence is weak		
Llama: best-depth ridge	-0.180 [2/10 pos.]	-0.238 [1/4 pos.]
Llama: warped-path z	-1.69 [9/10 neg.]	-1.69 [3/4 neg.]
DINOv2: best-depth ridge	-0.063 [5/10 pos.]	$+0.213$ [3/4 pos.]
DINOv2: warped-path z	$+0.50$ [2/10 neg.]	-0.84 [3/4 neg.]
3. Category structure resolves coarse-to-fine		
Coarse $\rightarrow$ fine onset	65 $\rightarrow$ 85 ms	65 $\rightarrow$ 125 ms
4. Controls bound the per-slice component		
Gaussian CCA ρ_1: category / Llama / DINOv2	0.92 / 0.95 / 0.97	0.93 / 0.96 / 0.95
Rank vs. null: Llama	0.225 vs. 0.079	0.271 vs. 0.066
Rank vs. null: DINOv2	0.369 vs. 0.081	0.291 vs. 0.069

apply our method to EEG and MEG responses to natural images, and to two foundation models chosen to differ in input modality: Llama-3-8B and DINOv2 ViT-L/14. Brain dynamics are measured using overlapping time windows and foundation model dynamics using layer depth. Our contributions are both methodological and empirical. On the methodological side, we propose how covariance of stimulus responses can be used to generate an SPD manifold on which relationships between stimuli can be defined to construct and measure representational trajectories. Our empirical findings are summarized in Table 1. Figure 1 outlines the construction of the paper.

2 Related Work

The analysis of covariance matrices on the SPD manifold is well developed: rate-invariant frameworks represent covariance trajectories by transported square-root vector fields and compare them by geodesic distance in a reparameterization-invariant quotient (Zhang et al., 2018). Our velocity and acceleration are local instances of this machinery.Dynamical Similarity Analysis (Ostrow et al., 2023) argues that systems can share *dynamics* without sharing *geometry*, and implements this by fitting linear Koopman operators and comparing them with a vector-field-aware Procrustes metric. Our measure is complementary: rather than fitting a global operator, it compares the basis-free *relational angular geometry* of local velocity fields, yielding an alignment-free statistic.

Riemannian geometry is commonly used in modern EEG/MEG analysis (Yger et al., 2017; Sabbagh et al., 2020). For example, Barachant et al. (2012) showed that AIRM-based classification of single-trial covariances outperforms a Euclidean approach.

Recent literature relates LLM representations to brain responses, predominantly through linear encoding from contextual embeddings (Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Toneva & Wehbe, 2019), improving with model scale and instruction tuning (Aw et al., 2024). We depart from the encoding-model paradigm: rather than fitting a map from model features

to neural responses, we compare the two systems' representational geometry without any learned alignment. Closer to our findings, Caucheteux et al. (2023) report a hierarchical correspondence between LLM depth and cortical regions integrating over different timescales, and Raugel et al. (2025) find that deeper layers track later brain responses across 17 LLMs.

In the visual domain, Doerig et al. (2025) report substantial LLM-text-embedding correspondence with high-level visual responses, while Conwell et al. (2024) show that many vision models cluster near the noise ceiling after voxelwise encoding fits, despite architectural differences. More directly, Rong et al. (2025) build *encoding* models on large-scale EEG combining a vision DNN and an LLM, and report that the vision network explains earlier, broadband responses while the LLM explains later, low-frequency semantic activity, with the fusion outperforming either alone.

Category structure in cortical responses is well studied. Cichy et al. (2014) report that individual images are discriminated early in MEG while ordinate and superordinate category levels emerge later; Khaligh-Razavi & Kriegeskorte (2014) model the categorical component explicitly, fitting category-cluster predictors to each computational model, and find that supervised networks still require category remixing to explain inferotemporal cortex; and Sexton & Love (2022) show that under a stricter direct-interface criterion, all ventral-stream regions correspond best with *later* layers indicating every processing stage carries higher-level category information.

EEG and MEG have long been jointly analyzed for source localization, where their complementary sensitivities can improve source estimates (Sharon et al., 2007; Molins et al., 2008) and their signal-to-noise ratios for cortical sources have been compared (Goldenholz et al., 2009). Those studies ask where signals originate; we use the two modalities as independent measurements.

3 METHOD

3.1 DATA

We use two publicly available datasets built on the THINGS image set (Hebart et al., 2019). THINGS-EEG2 (Gifford et al., 2022) provides 64-channel EEG from ten participants; its test partition contains 200 images at 80 repetitions each, which we trial-average within image to obtain high-signal single-image responses. THINGS-MEG (Hebart et al., 2023) provides CTF MEG from four participants with 200 test images at 12 repetitions each. We concentrate on the occipito-temporal channels in our analysis. For the foundation models, we use Llama-3-8B (Grattafiori et al., 2024) for the text model and DINOv2 ViT-L/14 (Oquab et al., 2024) as the image counterpart.

3.2 NEIGHBORHOOD COVARIANCES

For each image I, we define its semantic neighborhood $\mathcal{N}_k(I)$ as its $k = 20$ nearest images under Euclidean distance in CLIP image-embedding space (Radford et al., 2021). For brain modality m, subject s, and sliding window centered at time t (width 50 ms, stride 20 ms), let $\bar{\mathbf{r}}_m^{(s)}(J,t)$ be the window-averaged response to image J. The brain neighborhood covariance is

$$C_m^{(s)}(I,t) \;=\; \mathrm{LW}\Big[\mathrm{Cov}_{J\in\mathcal{N}_k(I)}\big\{\bar{\mathbf{r}}_m^{(s)}(J,t)\big\}\Big] + \epsilon\,\mathbf{I}, \tag{1}$$

where LW$[\cdot]$ denotes Ledoit–Wolf shrinkage (Ledoit & Wolf, 2004) and $\epsilon = 10^{-6}$. For Llama-3-8B, each image is presented to the model as its THINGS object label (the concept name, e.g. "aircraft carrier"), with no caption or prompt template; the layer-ℓ neighborhood covariance $C_L(I,\ell)$ is defined analogously from mean-pooled hidden states. All covariances are projected to $\mathrm{SPD}(d{=}8)$ by per-system PCA.

3.3 TANGENT-SPACE REPRESENTATION

The AIRM inner product (Pennec et al., 2005) at a base point C is $\langle U,V\rangle_C = \mathrm{Tr}(C^{-1}UC^{-1}V)$. We consider the tangent spaces of each system at its *own* Fréchet mean $\bar{C}$,

$$X(I,\cdot) \;=\; \log_{\bar{C}}\, C(I,\cdot), \tag{2}$$

whitened so the AIRM inner product becomes Euclidean and half-vectorized (off-diagonals weighted by $\sqrt{2}$) to a vector in $\mathbb{R}^{d(d+1)/2}$. A single base point is used for each system's whole trajectory (for the brain, per subject).

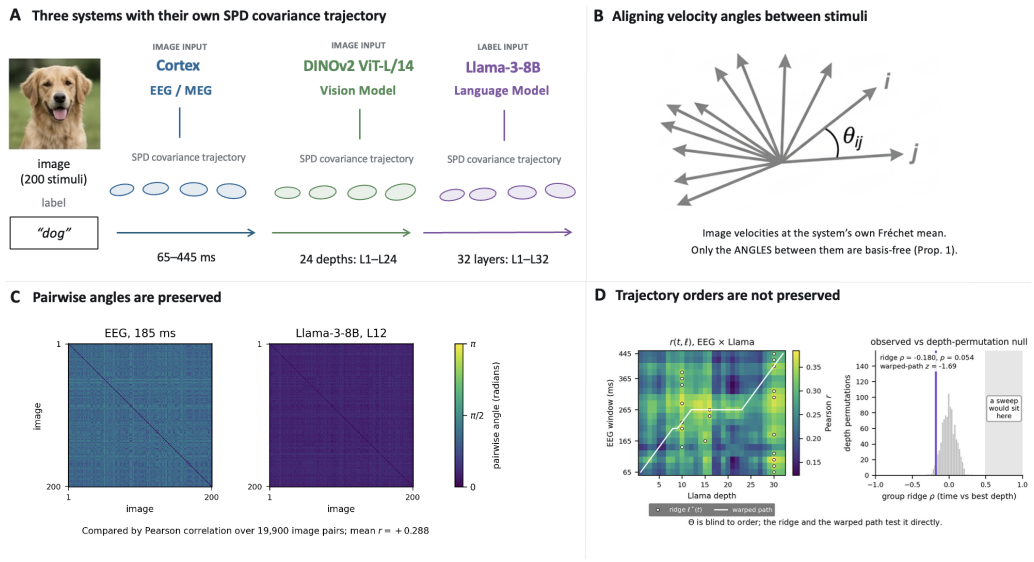


Figure 1: Construction overview. **(A)** Shared stimuli are presented as images to cortex and DINOv2 and as object labels to Llama. Each system yields a trajectory of neighborhood covariances on the SPD manifold, indexed by cortical time or model depth. **(B)** Covariances are mapped to the tangent space at each system's *own* Fréchet mean, and unit-normalized velocity directions are estimated. **(C)** Pairwise velocity angles form an image×image matrix Θ per slice. After within-system standardization (Eq. 6), matrices are compared using Eq. 7. **(D)** Order-sensitive statistics are computed from the resulting window×depth grid: the ridge tracks the best-matching depth, while dynamic time warping finds the least-cost monotone path (Eq. 8). Both are evaluated against model-depth permutations.

3.4 Velocity and Acceleration

The Riemannian velocity of the brain trajectory is the intrinsic time-derivative,

$$V_m^{(s)}(I,t) \;=\; \tfrac{1}{\delta}\,\log_{C_m^{(s)}(I,t)}\!\big(C_m^{(s)}(I,t{+}\delta)\big), \tag{3}$$

transported within the brain system to $\bar{C}$ and, for the angular analysis, unit-normalized to a *velocity direction*. Because consecutive velocities live in different tangent spaces, acceleration is the *covariant* difference: $V_{t+\delta}$ is parallel-transported back to C_t before subtraction,

$$A_m^{(s)}(I,t) \;=\; \tfrac{1}{\delta}\big[\,P_{C_{t+\delta}\to C_t}\,V_m^{(s)}(I,t{+}\delta) \;-\; V_m^{(s)}(I,t)\,\big], \tag{4}$$

and then transported to $\bar{C}$ like the velocity. Equations 3 and 4 state the intrinsic construction. The implementation evaluates them in a single tangent chart at the grand Fréchet mean rather than composing per-point logarithms with parallel transport, which is accurate here because the trajectories stay near $\bar{C}$ relative to the curvature scale: the component correlation between the chart-based and covariant velocity directions are close to one across the four trajectory types (EEG 0.999, MEG 0.999, Llama 1.000, DINOv2 0.997).

We fit a local quadratic over five consecutive slices (the nearest five at either end), examine its first derivative (a Savitzky–Golay filter of window five and polynomial order two), and then unit-normalize. The model's trajectory unfolds across *depth* rather than time, so the LLM velocity is the layer-to-layer analog $V_L(I,\ell) = \frac{1}{\Delta\ell}\log_{C_L(I,\ell)}(C_L(I,\ell{+}\Delta\ell))$.

3.5 Angular Concordance

To compare motion geometry across systems without a common frame, we compute, for each system slice (a brain window or an LLM layer), the image×image matrix of angular distances between per-image velocity directions,

$$\theta_{ij} \;=\; \arccos\frac{\langle x_i, x_j\rangle}{\|x_i\|\,\|x_j\|} \in [0,\pi], \tag{5}$$

Writing $\Theta_s \in \mathbb{R}^{N\times N}$ for the matrix $[\theta_{ij}]$ at slice s, we first put each system on a common scale. Angles are standardized within a system, over the off-diagonal entries of its whole slice stack,

$$\widetilde{\Theta}_s \;=\; \frac{\Theta_s - \mu}{\sigma}, \qquad \mu = \operatorname*{mean}_{s,\, i\neq j} \theta_{ij}^{(s)}, \quad \sigma = \operatorname*{sd}_{s,\, i\neq j} \theta_{ij}^{(s)}, \tag{6}$$

separately for each subject on the brain side and once for each model. The diagonal is excluded throughout: Θ_s is symmetric with zero diagonal.

Two standardized slices are then compared by the correlation of their angle vectors,

$$r(t,\ell) \;=\; \operatorname{corr}\big(\widetilde{\Theta}^B_t,\, \widetilde{\Theta}^L_\ell\big) \qquad \text{over the } N(N-1) \text{ off-diagonal entries.} \tag{7}$$

Because Eq. 6 applies one affine map to each system as a whole, it leaves r unchanged.

A pairwise angle is invariant to any rotation of a system's frame, so the comparison requires no common basis and no learned map: it asks only whether the two systems assign the same relative angles to the same images. The zero of r is calibrated empirically by an image-permutation level, obtained by shuffling the model's images before Θ is built, which removes the correspondence and leaves the marginal angle distribution intact. Graph-null comparisons (Sec. 4.3, App. J, Table 1, Figure 7C) use the rank (Spearman) version of Eq. 7 throughout.

3.6 An order-sensitive statistic

To ask whether the correspondence is *ordered* across depth, we compute, per brain window t, the best-matching layer $\ell^*(t) = \arg\max_\ell r(t,\ell)$ and test the rank correlation between window time and ℓ^* (the "ridge"). The ridge requires the correspondence to be *rate-matched* in the sense that ℓ^* must rise steadily. Our second test is monotone but rate-free. Treating $1 - r(t,\ell)$ as a cost, dynamic time warping (DTWD) (Sakoe & Chiba, 1978) finds the monotone path through the window×depth grid that minimizes the total cost, allowing either axis to advance while the other waits, and we report the mean cost along that path,

$$\mathrm{DTWD} \;=\; \frac{1}{|\mathcal{P}^*|} \sum_{(t,\ell)\in\mathcal{P}^*} \big(1 - r(t,\ell)\big), \qquad \mathcal{P}^* = \operatorname*{argmin}_{\text{monotone } \mathcal{P}} \sum_{(t,\ell)\in\mathcal{P}} \big(1 - r(t,\ell)\big), \tag{8}$$

normalized by path length so that paths of different lengths are comparable. Warping by repetition allows both systems to traverse their axes without skipping, so a path may pause but may not omit a window or a depth. The two axes are left at their native resolutions throughout, the brain at 20 windows and the model at its full depth, 32 layers for Llama and 24 for DINOv2. Both statistics are therefore assessed against the same null of orderings of the model's own depth axis. Both are defined per subject, so both are permuted per subject.

3.7 Cross-validated Gaussian-CCA dependence

Since our descriptors are high-dimensional relative to the number of images, we quantify dependence with a canonical-correlation statistic rather than a density- or neighbor-based estimator. For canonical correlations $\{\rho_i\}$, we report the *Gaussian-equivalent dependence*

$$S(X;Y) \;=\; -\tfrac{1}{2} \sum_i \log_2\!\big(1 - \rho_i^2\big), \tag{9}$$

in bits. Under joint Gaussianity S equals the mutual information. The estimator is cross-validated, and the ρ_i in Eq. 9 is measured on the held-out partition. Conditional information $I(X;Y \mid Z)$ is obtained by ridge-residualizing both X and Y on Z within the training partition before the canonical-correlation step. Two conditionings are used: on speed, to ask whether velocity *direction* is informative beyond how fast the representation moves (with speed and direction obtained by factorizing the velocity into magnitude and unit direction); and on category, to ask whether anything survives the category structure of Sec. 4.3. We report every value as an *excess* over an image-permutation null with 200 shuffles (50 for the random-feature controls of App. D), which permutes the pairing of a model row with a brain row. Quoted canonical correlations are the held-out first, ρ_1.

3.8 BASIS INVARIANCE

In order to understand whether a comparison is "basis-free", we write the AIRM inner product at a base point P as $\langle U, W\rangle_P = \mathrm{Tr}(P^{-1}UP^{-1}W)$, and let $u_i \in T_{\bar{C}}\mathrm{SPD}(d)$ be the tangent velocity of image i at the Fréchet mean $\bar{C}$: either the time-derivative of the tangentized state $u_i = \frac{d}{dt}\log_{\bar{C}}(C_i(t))$, or the local velocity $\log_{C_i}(D_i)$ parallel-transported to $\bar{C}$. The pairwise angle is $\theta(u_i, u_j) = \arccos\big(\langle u_i, u_j\rangle_{\bar{C}}/(\|u_i\|_{\bar{C}}\|u_j\|_{\bar{C}})\big)$, and this equals the whitened-vech angle of Sec. 3.3, since $\langle U, W\rangle_{\bar{C}} = \langle \bar{C}^{-1/2}U\bar{C}^{-1/2},\ \bar{C}^{-1/2}W\bar{C}^{-1/2}\rangle_F$ and $\mathrm{vech}_{\sqrt{2}}$ is a Frobenius isometry. The following proposition (proof in App. A) shows that the angular concordance is preserved under a congruence transform.

Proposition 1. *For any $A \in GL(d)$, transform every point of every trajectory by the congruence $C_i \mapsto C_i' = AC_iA^\top$. Then for all i, j, $\theta(u_i', u_j') = \theta(u_i, u_j)$.*

Proposition 2 in App. A extends this to per-image rescaling, $C_i(t) \mapsto c_i\, C_i(t)$ with $c_i > 0$ constant along the trajectory, so the statistic reads the shape of representational change and not the magnitude of the response to any one image.

4 RESULTS

4.1 THE ANGULAR GEOMETRY OF THE MOTION IS SHARED

Here, we measure the angle between the velocity directions of each pair of images in the cortex (Sections 3.4 and 3.5) and compare that with similar angles in the two foundation models. Collecting those angles over all $19{,}900$ image pairs gives one matrix Θ per slice per system; the entries are standardized within each system and the two are compared through the correlation of Eq. 7 (Prop. 1).

Cortex and the two foundation models assign markedly similar relative angles. Against Llama-3-8B the correlation is $+0.2632$ in EEG and $+0.2835$ in MEG; against DINOv2 it is $+0.3913$ and $+0.3189$. All four are positive in every subject ($10/10$, $4/4$, $10/10$, $4/4$) and in every cell of the group-mean window×depth grid, the smallest cell of the four being $+0.1007$. Per-subject values run 0.2214–0.2917 and 0.3701–0.4110 in EEG and 0.1936–0.3393 and 0.2125–0.3647 in MEG, the wider MEG spread reflecting four subjects rather than ten. Shuffling the model's images before the angles are formed removes the effect entirely, leaving -0.0003, -0.0001, -0.0002 and -0.0002.

The two models reach their closest agreement at different depths (Figure 2). In EEG, DINOv2, which receives pixels, climbs from 0.3135 at its first block to a peak at 48% of the network (0.4542), ending at 0.3767. Llama, handed the object label, has a non-monotone profile whose highest point lies at 94% of depth (0.3315, against 0.2253 at the first block). MEG does not reproduce these peak locations: Llama peaks at 35% of depth and DINOv2 at 65%.

Repeating the identical statistic one derivative up, on acceleration rather than velocity directions, raises it by 2–10 per cent in all four system pairs (0.288 and 0.406 in EEG, 0.289 and 0.334 in MEG, positive in every subject).

4.2 ORDER-PRESERVING CORRESPONDENCE IS RATE-FREE, NOT RATE-MATCHED

Here, we ask whether the two systems visit the shared geometry in the same order: whether early cortical windows correspond to shallow model layers and late windows to deep ones. Section 3.6 gives two tests that differ in what they demand. The rate-free test finds an ordering. Under dynamic time warping (Eq. 8), the warped path against Llama is shorter than under permuted depth orderings in 9/10 EEG subjects and 3/4 MEG subjects, with mean $z = -1.69$ in both cohorts (7 of 14 subjects clear their own depth-permutation null at $p < 0.05$, against 0.7 expected by chance). Against DINOv2 it does not: $+0.50$ in EEG, with only 2/10 subjects below zero, and -0.84 in MEG with 3/4 (1 of 14 subjects significant, against 0.7 expected).

The rate-matched test finds no ordering for either model: the ridge is -0.180 against Llama in EEG and -0.238 in MEG, -0.063 and $+0.213$ against DINOv2, and none is significantly positive (App. H). The two tests are not in conflict. The ridge requires the best-matching depth to advance steadily, while Eq. 8 requires only a cheap monotone path through the window×depth grid. Against Llama

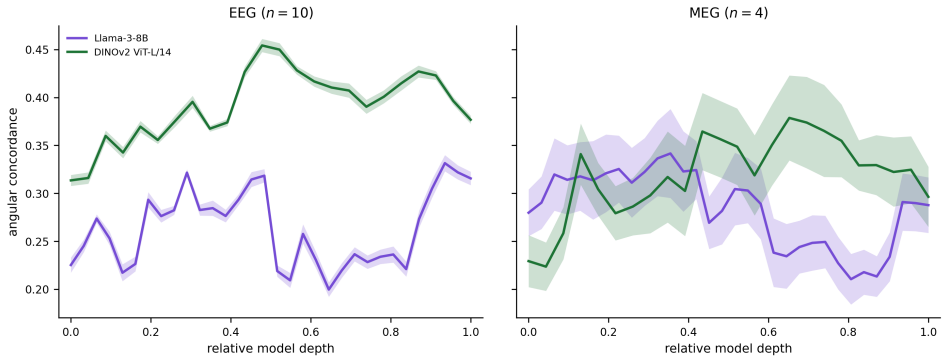


Figure 2: Concordance against relative model depth, both cohorts on a shared vertical axis. Left, EEG (n=10); right, MEG (n=4). In EEG the two models separate: DINOv2 peaks at 48% of depth, while Llama's non-monotone profile peaks 94%. In MEG the peaks lie at 35% for Llama and 65% for DINOv2, so peak locations do not replicate across cohorts. Shading is s.e.m. over subjects.

the grid admits such a path in almost every subject while the best-matching depth wanders, so the ordering lives in the correlation profile as a whole rather than any single best layer.

The model that carries the ordering is the label-fed one. This is the reverse of the asymmetry a feedforward account predicts: on that account DINOv2, which receives pixels and must construct category-level geometry across its depth, would offer cortex a stage order to follow, while Llama, which receives the object label, would not.

4.3 Category explains only part of the similarity

Here, we ask the reasons behind the shared geometry: whether the two systems agree only about which images belong to the same named class or because of the CLIP neighborhood over which both pipelines are built (CLIP's block structure is itself largely category co-membership; same-category image pairs share 8.2 of 20 neighbors on average against 2.2 for different-category pairs).

For class membership under cross-validated Gaussian-CCA, a one-hot category model reaches $\rho_1 = 0.92$ in EEG and 0.93 in MEG, below the real brain↔Llama velocity coupling (0.95, 0.96) and the brain↔DINOv2 coupling (0.97, 0.95) in every subject: at the level of statistical *dependence*, knowing the model tells more than knowing the category. The *angular* statistic behaves differently. Partialling out a one-hot category RDM removes only a minority: 0.263 falls to 0.189 against Llama and 0.391 to 0.327 against DINOv2 in EEG, and MEG retains a similar share (85% and 84%). Because 72% to 85% survives, the label is not most of what the two systems agree about.

The CLIP neighborhood graph has a stronger influence. Replacing both systems' features with independent Gaussian noise while keeping the real neighborhoods reproduces 0.079 of the observed 0.225 in EEG and 0.066 of 0.271 in MEG, both scored with the same rank concordance, 35% and 24% of the statistic, before either system's own organization enters. Taken together, the above analysis means that the correspondence at the category level is graded. What the two systems share is not the partition into named classes, but the neighborhood structure.

4.4 Velocity concordance is affected by category membership

Section 4.3 bounds per-slice concordance against a category model and stimulus-graph null. Here we ask the complementary question: at what grain does the correspondence live, and where along model depth does each grain sit? The $N(N-1)/2$ image pairs are split into within- and across-category sets, each restricted before the correlation is computed, so the statistic is recomputed within grain. Each cohort uses its own 200-image test set and taxonomy: 20 EEG categories yield 2083 within- and 17817 across-category pairs, while 64 MEG categories yield 1188 and 18712, respectively.

Within-category concordance exceeds across-category for DINOv2 and Llama in EEG (0.445 vs. 0.313; 0.439 vs. 0.152) and MEG (0.431 vs. 0.256; 0.330 vs. 0.235; Fig. 3).All peaks are positive.

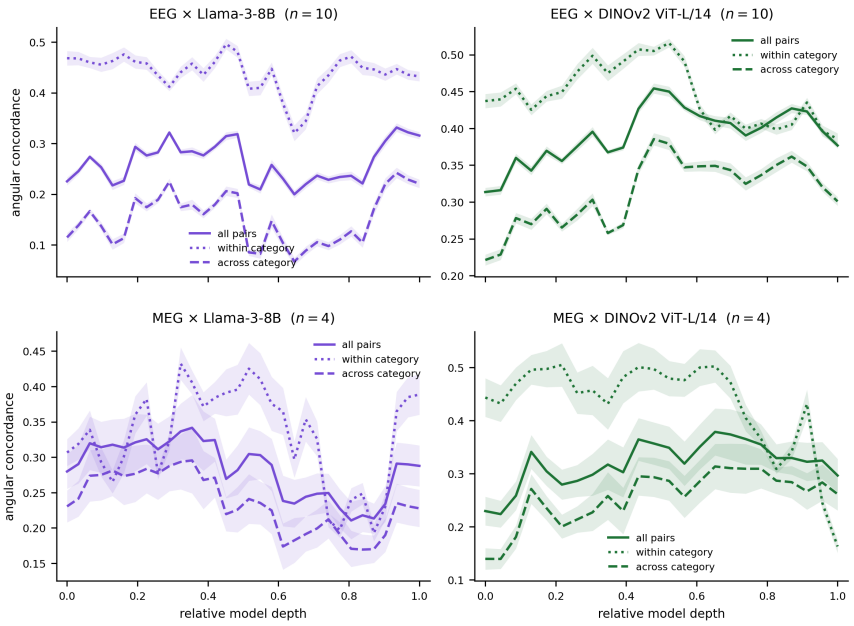


Figure 3: Depth profile split by within- and across-category pairs, each cohort on its own test set and taxonomy. Correlations are recomputed within the restricted pair set rather than inherited from the full triangle. Within-category pairs (dotted) exceed all pairs (solid), which exceed across-category pairs (dashed), at most depths in every panel (72–100% of depths). Shading is s.e.m. over subjects.

Two controls exclude these artifacts. Subsampling across-category pairs to match the within-category count over twenty draws leaves levels unchanged in EEG (0.314, 0.152) and MEG (0.255, 0.233), ruling out pair-count imbalance. Removing the two largest categories, which supply 77% of within-category pairs, preserves this advantage on average across cohorts and models (App. C).

4.5 Category structure resolves coarse-to-fine, on different timelines

Hierarchical clustering of category-separation dynamics resolves coarse before fine distinctions (Figure 4): $65 \rightarrow 85$ ms in EEG and $65 \rightarrow 125$ ms in MEG. Medians differ by $+30$ ms in EEG ($6/10$ positive, four tied; $p = 0.031$) and $+65$ ms in MEG ($3/4$ positive, one tied; $p = 0.250$), though four subjects cannot resolve the latter. Both models resolve coarse before fine over depth; at its first sampled layer, the label-fed model resolves more splits (60% versus 42% on EEG images). Onset correlates with split image count ($\rho = -0.33$ in EEG), but the ordering holds in the sole quartile containing both split types ($+60$ ms), indicating an ordinal tendency rather than a strict latency.

5 Discussion

We introduced an alignment-free framework for comparing representational trajectories across independent systems without learning a cross-system map. Each system remains in its own representational space: nearby-stimulus responses define covariance trajectories on the SPD manifold, velocities describe change along the system-specific axis, and systems are compared through the pairwise angular geometry of those velocity directions. The resulting statistic is invariant to independent invertible transformations of each system's covariance coordinates and to per-image rescaling, separating representational correspondence from the choice of basis or feature units.

Previous brain–LLM work reports hierarchical correspondences for continuous language: Caucheteux et al. (2023) relate LLM depth and prediction horizon to a hierarchy of cortical regions, and Raugel et al. (2025) find deeper layers tracking later responses across seventeen models. Both use minutes of naturalistic speech, where deeper layers encode extended linguistic context. We

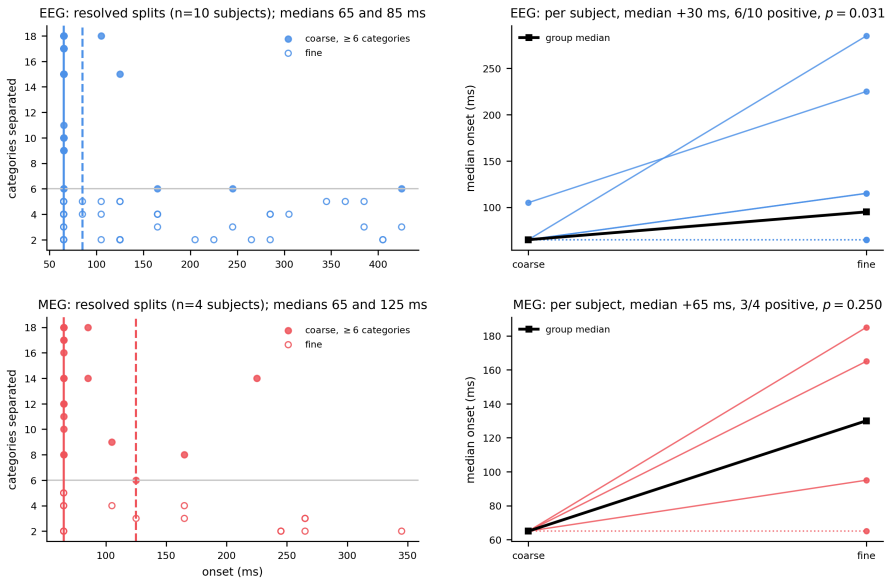


Figure 4: Category structure resolves coarse to fine within a single occipito-temporal SPD. Node onset is the first of two consecutive windows in which between- minus within-group angular separation exceeds a permutation threshold; unresolved nodes are omitted, and splits separating at least six categories are coarse. Left: pooled node onsets versus categories separated; gray marks the coarse cutoff and solid/dashed lines the coarse/fine medians. Right: paired coarse/fine medians per subject; black marks group medians and dotted lines ties.

find an ordered correspondence between cortical time and Llama depth from a static image, a 475 ms epoch, and an isolated object label with no context to integrate. Depth-to-time ordering therefore does not require extended context to accumulate. It does require the rate-free test: the ordering is absent from the best-matching-layer trajectory in the same data.

A second comparison concerns which model stage best matches a cortex response. Sexton & Love (2022) find all ventral-stream regions correspond best with *later* layers under a strict direct-interface criterion. Our DINOv2 result matches this ordering only loosely: correspondence peaks at mid-depth in EEG (48%) and at 65% in MEG. The language model shows the pattern more clearly, peaking at 94% in EEG, though this does not replicate in MEG (35%). Separately, Conwell et al. (2024) find many vision models near the noise ceiling despite architectural differences.

The third comparison is to work that treats category structure as the object rather than the nuisance. Consistent with Khaligh-Razavi & Kriegeskorte (2014), the identically scored one-hot category model falls slightly below the real brain↔model coupling in both cohorts.

Against standard alternatives, the framework offers scope rather than score (App. L): orthogonal Procrustes fits a global map with no image-pair term. CKA's global centering prevents recomputing within a subset of pairs. RSA supports the same pair-restricted analyses but compares state geometry rather than motion, and is not used to ask the order question. Finally, DSA must fit stable dynamics along each axis, whereas ours needs only monotone traversal without fitting a cross-system map.

Geometric and order-preserving correspondence are not two readings of one quantity: across our two models they run in opposite directions. DINOv2 shares more angular geometry with cortex (0.3913 against 0.2632 in EEG, per-subject ranges disjoint) while Llama carries the monotone ordering (9/10 EEG, 3/4 MEG). A single alignment score would rank the two models in whichever direction it was sensitive to, without revealing that the ranking was a choice.

Our framework is meant to be general. Although demonstrated on EEG, MEG, Llama, and DINOv2, it requires only shared stimuli and sequential representations. The axis may be cortical time, model depth, optimization, recurrent iteration, diffusion denoising, or another system-specific progression. Within-system angular relations avoid cross-system dimensional matching and encoding models.

AI USE STATEMENT

We used generative AI tools as a coding and writing assistant: to help implement the analysis and figure code, and to draft and edit prose. All experiments were run and their outputs inspected by the authors, all numerical claims were checked against the saved outputs, and the authors take full responsibility for the content, claims, and artifacts of this work.

ETHICS STATEMENT

This work analyses two publicly available human neuroimaging datasets, THINGS-EEG2 (Gifford et al., 2022) and THINGS-MEG (Hebart et al., 2023). Both were collected and publicly released by their original authors with the participants' informed consent and the approval of the responsible ethics committees, as documented in those publications; we collected no new data. The two foundation models are used under their published licenses. We are not aware of further ethical concerns raised by this work.

REPRODUCIBILITY STATEMENT

Section 3 and Appendices A–L describe the analysis and its settings: the image neighborhoods, the PCA reductions and neighborhood covariances (Eq. 1), the tangent chart at each system's Fréchet mean (Sec. 3.3), the derivative filter (Sec. 3.4), the concordance and order statistics (Secs. 3.5–3.6), the dependence measure (Sec. 3.7), and the permutation and sign-flip tests used for inference. Proposition 1 is proved in App. A. Both datasets and both foundation models are publicly available.

REFERENCES

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. Instruction-tuning aligns LLMs to the human brain. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=nXNN0x4wbl.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012. doi: 10.1109/TBME.2011.2172210.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, 2022. doi: 10.1038/s42003-022-03036-1.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441, 2023. doi: 10.1038/s41562-022-01516-2.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, 2014. doi: 10.1038/nn.3635.

Alex Clarke, Kirsten I. Taylor, Barry Devereux, Billi Randall, and Lorraine K. Tyler. From perception to conception: How meaningful objects are processed over time. *Cerebral Cortex*, 23(1):187–197, 01 2013. ISSN 1047-3211. doi: 10.1093/cercor/bhs002. URL https://doi.org/10.1093/cercor/bhs002.

Alex Clarke, Barry J. Devereux, Billi Randall, and Lorraine K. Tyler. Predicting the time course of individual objects with MEG. *Cerebral Cortex*, 25(10):3602–3612, 2015. doi: 10.1093/cercor/bhu203.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, 15:9383, 2024. doi: 10.1038/s41467-024-53147-y.

Adrien Doerig, Tim C. Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8):1220–1234, 2025. doi: 10.1038/s42256-025-01072-0.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119754. URL https://www.sciencedirect.com/science/article/pii/S1053811922008758.

Daniel M. Goldenholz, Seppo P. Ahlfors, Matti S. Hämäläinen, Dahlia Sharon, Mamiko Ishitobi, Lucia M. Vaina, and Steven M. Stufflebeam. Mapping the signal-to-noise-ratios of cortical sources in magnetoencephalography and electroencephalography. *Human Brain Mapping*, 30 (4):1077–1086, 2009. doi: 10.1002/hbm.20571. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.20571.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3): 369–380, 2022. doi: 10.1038/s41593-022-01026-4. URL https://doi.org/10.1038/s41593-022-01026-4.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, A. Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv*, 2024.

Martin N.. Hebart, Adam H. Dickter, Alexis. Kidder, Wan Y. Kwok, Ana. Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019. doi: 10.1371/journal.pone.0223792.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. doi: 10.7554/eLife.82580.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014. doi: 10.1371/journal.pcbi.1003915.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

A. Molins, S. M. Stufflebeam, E. N. Brown, and M. S. Hämäläinen. Quantification of the benefit from integrating MEG and EEG data in minimum ℓ_2-norm estimation. *NeuroImage*, 42(3):1069–1077, 2008. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2008.05.064. URL https://www.sciencedirect.com/science/article/pii/S1053811908007143.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. doi: 10.48550/arXiv.2304.07193. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Mitchell Ostrow, Adam Eisen, Leo Kozachkov, and Ila Fiete. Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 33824–33837. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1467. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6ac807c9b296964409b277369e55621a-Paper-Conference.pdf.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2005. doi: 10.1007/s11263-005-3222-z.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Joséphine Raugel, Jérémy Rapin, Stéphane d'Ascoli, Valentin Wyart, and Jean-Remi King. Scaling and context steer llms along the same computational path as the human brain. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 53588–53616. Curran Associates, Inc., 2025. doi: 10.52202/085713-1787. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4d61054c13fb29ef36028f18353df40e-Paper-Conference.pdf.

Boyan Rong, Alessandro Thomas Gifford, Emrah Düzel, and Radoslaw Martin Cichy. The time course of visuo-semantic representations in the human brain is captured by combining vision and language models. *arXiv preprint arXiv:2506.19497*, 2025. doi: 10.7554/elife.108915.

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A. Engemann. Predictive regression modeling with meg/eeg: from source power to signals and cognitive states. *NeuroImage*, 222:116893, 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.116893. URL https://www.sciencedirect.com/science/article/pii/S1053811920303797.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978. doi: 10.1109/TASSP.1978.1163055.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: doi:10.1073/pnas.2105646118.

Nicholas J Sexton and Bradley C Love. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, 8(28):eabm2219, 2022. doi: 10.1126/sciadv.abm2219.

Dahlia Sharon, Matti S Hämäläinen, Roger BH Tootell, Eric Halgren, and John W Belliveau. The advantage of combining MEG and EEG: comparison to fmri in focally stimulated visual cortex. *NeuroImage*, 36(4):1225–1235, 2007. doi: 10.1016/j.neuroimage.2007.03.066.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Neural Information Processing Systems*, 2019.

Florian Yger, Maxime Berar, and Fabien Lotte. Riemannian approaches in brain-computer interfaces: a review. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10): 1753–1762, 2017. doi: 10.1109/tnsre.2016.2627016.

Zhengwu Zhang, Jingyong Su, Eric Klassen, Huiling Le, and Anuj Srivastava. Rate-invariant analysis of covariance trajectories. *Journal of Mathematical Imaging and Vision*, 60(8):1306–1323, 2018. doi: 10.1007/s10851-018-0814-0.

Appendix: Table of contents

	Appendix	What it adds	Supports
A	Proof of the congruence-invariance proposition	The eight-step proof of Prop. 1, and why the invariance stops at the reduced covariances rather than reaching raw sensors.	§3.8
B	The depth profile at the stimulus level	The pooled curve resolved per image: broadly distributed peaks for Llama, mid-depth concentration for DINOv2, and why peak location is the weaker quantity.	§4.1
C	Category grain: peak locations and order test	Where each grain peaks along depth, why the within-category peak reflects two dominant categories, and the order test restricted to each grain.	§4.4
D	Reliability of velocity and acceleration	Split-half reliability of velocity and acceleration in both cohorts, model depth trends, and angular-matrix reliability.	§4.1
E	The angular geometry of the motion is preserved across systems	Per-cell angular concordance and the full-depth comparison showing higher concordance for DINOv2 than Llama.	§4.1
F	Comparing recording modalities with the foundation models.	EEG↔MEG concordance on the 28 shared concepts, and how it compares with each instrument's agreement with Llama and DINOv2.	§4.1
G	Effect of pooling the slices	Two pooled implementations, warped versus fixed alignment, and controls separating slice matching from sequence order.	§4.1
H	The rate-matched order test.	Ridge-based ordering results and depth-permutation tests, with reference to the rate-free comparison in §4.2.	§4.2
I	The shared structure goes beyond category	One-hot category comparisons, residual concordance after controlling for category, and the contribution of shared neighborhoods.	§4.3
J	The graph-scaffold null	Why a shared neighborhood graph induces concordance on its own, and the full set of nulls from the independent-graph floor to the observed value.	§4.3
K	Separation order of categories	Category-separation onsets in EEG and MEG, the vision-model comparison, and the image-count confound.	§4.5
L	Comparison against standard methods	Conceptual comparison with RSA, CKA, Procrustes and DSA, focusing on supported analyses and fitting assumptions.	§3.5

A PROOF OF THE CONGRUENCE-INVARIANCE PROPOSITION

Proposition 1. *For any $A \in GL(d)$, transform every point of every trajectory by the congruence $C_i \mapsto C_i' = AC_iA^\top$. Then for all i, j, $\theta(u_i', u_j') = \theta(u_i, u_j)$.*

Proof. Write $\varphi_A(C) = ACA^\top$ and apply it to every point of every image's trajectory in one system, $C_i(t) \mapsto C_i'(t) = AC_i(t)A^\top$, with the *same* A for all images and all slices; primes denote quantities computed from the transformed data. Recall $\langle U, W\rangle_P = \mathrm{Tr}(P^{-1}UP^{-1}W)$ and $\delta(P,Q) = \|\log(P^{-1/2}QP^{-1/2})\|_F$. Step 1 ($\varphi_A$ is a bijection of $\mathrm{SPD}(d)$).
$ACA^\top$ is symmetric, and for $x \neq 0$ we have $x^\top ACA^\top x = (A^\top x)^\top C(A^\top x) > 0$ because $A^\top x \neq 0$. So φ_A maps $\mathrm{SPD}(d)$ into itself, and $\varphi_{A^{-1}}$ inverts it. Step 2 (φ_A is an isometry).
Since $P^{-1/2}QP^{-1/2}$ is similar to $P^{-1}Q$, the distance is spectral, $\delta(P,Q)^2 = \sum_k \log^2 \lambda_k(P^{-1}Q)$. Under the congruence,

$$(APA^\top)^{-1}(AQA^\top) = A^{-\top}P^{-1}A^{-1}AQA^\top = A^{-\top}\big(P^{-1}Q\big)A^\top, \tag{10}$$

a similarity transform of $P^{-1}Q$ by $A^\top$. Similar matrices share eigenvalues, so $\delta(APA^\top, AQA^\top) = \delta(P,Q)$. Step 3 (the Fréchet mean is equivariant).
Let $F(M) = \sum_i \delta(M, C_i)^2$, minimized at $\bar{C}$, and let $G(N) = \sum_i \delta(N, C_i')^2$ be the objective on the transformed data. By Step 2, $G(AMA^\top) = \sum_i \delta(AMA^\top, AC_iA^\top)^2 = F(M)$. Every N equals $AMA^\top$ for exactly one M (Step 1), so $G(N) = F(M) \geq F(\bar{C}) = G(A\bar{C}A^\top)$ for all N; hence $A\bar{C}A^\top$ minimizes G. The minimizer is unique, $\mathrm{SPD}(d)$ under AIRM being a Hadamard space on which F is strictly geodesically convex, so

$$\bar{C}' = A\bar{C}A^\top. \tag{11}$$

Step 4 (the logarithm is equivariant).
The geodesic from P to Q is $\gamma(t) = P^{1/2}(P^{-1/2}QP^{-1/2})^tP^{1/2}$, and $\log_P(Q) = \dot{\gamma}(0)$. An isometry carries geodesics to geodesics, so $t \mapsto A\gamma(t)A^\top$ is the geodesic from $APA^\top$ to $AQA^\top$; differentiating at $t = 0$, with A independent of t,

$$\log_{APA^\top}(AQA^\top) = A\,\log_P(Q)\,A^\top. \tag{12}$$

Step 5 (the velocity transforms by the same congruence).
For the tangentized-state construction $u_i = \frac{d}{dt}\log_{\bar{C}} C_i(t)$, Eqs. 11–12 give $\log_{\bar{C}'} C_i'(t) = A\log_{\bar{C}} C_i(t)A^\top$, and since A is constant in t differentiation passes through:

$$u_i' = Au_iA^\top. \tag{13}$$

For the transported-local-velocity construction the logarithm transforms by Eq. 12 and parallel transport commutes with isometries, the Levi-Civita connection being natural, so Eq. 13 holds again. Step 6 (the inner product is unchanged).
From Eq. 11, $(\bar{C}')^{-1} = A^{-\top}\bar{C}^{-1}A^{-1}$. Substituting this and Eq. 13, and canceling the three interior products $A^{-1}A = I$, $A^\top A^{-\top} = I$ and $A^{-1}A = I$,

$$\langle u_i', u_j'\rangle_{\bar{C}'} = \mathrm{Tr}\big(A^{-\top}\bar{C}^{-1}A^{-1}\,Au_iA^\top\,A^{-\top}\bar{C}^{-1}A^{-1}\,Au_jA^\top\big) = \mathrm{Tr}\big(A^{-\top}\,\bar{C}^{-1}u_i\bar{C}^{-1}u_j\,A^\top\big),$$

and cyclic invariance of the trace moves $A^\top$ to the front, cancelling $A^{-\top}$:

$$\langle u_i', u_j'\rangle_{\bar{C}'} = \mathrm{Tr}\big(\bar{C}^{-1}u_i\bar{C}^{-1}u_j\big) = \langle u_i, u_j\rangle_{\bar{C}}. \tag{14}$$

Step 7 (norms).
Setting $j = i$ in Eq. 14 and taking non-negative square roots gives $\|u_i'\|_{\bar{C}'} = \|u_i\|_{\bar{C}}$ for every i.
Step 8 (angles).
Numerator and both denominator factors are individually unchanged by Steps 6 and 7, so $\cos\theta(u_i', u_j') = \cos\theta(u_i, u_j)$; both angles lie in $[0, \pi]$, on which $\cos$ is injective, hence $\theta(u_i', u_j') = \theta(u_i, u_j)$. □

Note that every step is internal to one system, so the two systems may be transformed by different matrices A_1, A_2 at once: this is what makes the comparison basis-*free* rather than basis-*consistent*. The congruence acts on the already-reduced $d \times d$ covariances; principal-component analysis is not congruence-equivariant, since reducing $ACA^\top$ generally selects a different subspace than reducing C, so the invariance does not extend back through the reduction to raw sensors.

Proposition 2. *Fix a constant $c_i > 0$ for each image i and transform that image's whole trajectory by $C_i(t) \mapsto C_i'(t) = c_i\, C_i(t)$, the same c_i at every slice. Then $\langle u_i', u_j' \rangle_{\bar{C}'} = \langle u_i, u_j \rangle_{\bar{C}}$ for all i, j, and hence $\theta(u_i', u_j') = \theta(u_i, u_j)$.*

Proof. Step 1.
$c_i C = (\sqrt{c_i} I)\, C\, (\sqrt{c_i} I)^\top$, so each image is acted on by a congruence and therefore by an isometry, but the congruence differs across images, which is what Prop. 1 excludes. Step 2 (a scalar leaves the logarithm additively).
For $M \in \mathrm{SPD}(d)$ and $c > 0$, cI commutes with M, so $\log(cM) = (\log c) I + \log M$ and therefore, for any $P \in \mathrm{SPD}(d)$,

$$\log_P(cQ) = P^{1/2} \log\big(c\, P^{-1/2} Q P^{-1/2}\big) P^{1/2} = (\log c)\, P + \log_P(Q). \tag{15}$$

The log map is also positively homogeneous: substituting $(aP)^{-1/2}(aQ)(aP)^{-1/2} = P^{-1/2} Q P^{-1/2}$ into its definition gives

$$\log_{aP}(aQ) = a\, \log_P(Q), \qquad a > 0. \tag{16}$$

Step 3 (the mean scales by the geometric mean of the c_i).
The Fréchet mean over the n images and T slices satisfies the Karcher condition $\sum_{i,t} \log_{\bar{C}} C_i(t) = 0$. Put $\lambda = \big(\prod_i c_i\big)^{1/n}$ and write $c_i C_i(t) = \lambda \cdot (c_i/\lambda) C_i(t)$. Eqs. 16 and 15 give

$$\log_{\lambda \bar{C}}\big(c_i C_i(t)\big) = \lambda \Big[\log(c_i/\lambda)\, \bar{C} + \log_{\bar{C}} C_i(t)\Big]. \tag{17}$$

Summing over i and t, the first term vanishes because $\sum_i \log(c_i/\lambda) = 0$ by the definition of λ, and the second vanishes by the Karcher condition. So $\lambda \bar{C}$ satisfies the first-order condition on the transformed data, and by the uniqueness argument of Step 3 of Prop. 1,

$$\bar{C}' = \lambda\, \bar{C}. \tag{18}$$

Step 4 (every velocity scales by that one scalar).
In Eq. 17 the term $\lambda \log(c_i/\lambda) \bar{C}$ is constant in t, because c_i is constant along the trajectory, so it differentiates away:

$$u_i' = \tfrac{d}{dt} \log_{\bar{C}'} C_i'(t) = \lambda\, u_i. \tag{19}$$

For the transported-local-velocity construction the local velocity scales as $v_i' = c_i v_i$ by Eq. 16; parallel transport along the geodesic is $\Gamma_{P \to Q}(U) = (QP^{-1})^{1/2} U (QP^{-1})^{\top/2}$, and Eq. 18 gives $\big(\bar{C}'(C_i')^{-1}\big)^{1/2} = (\lambda/c_i)^{1/2} \big(\bar{C} C_i^{-1}\big)^{1/2}$, so the two factors combine as $(\lambda/c_i) \cdot c_i = \lambda$ and Eq. 19 holds again. The per-image scalars cancel; only their geometric mean survives, and it is common to all images. Step 5 (the metric absorbs the scalar).
$\langle U, W \rangle_{\lambda \bar{C}} = \mathrm{Tr}\big((\lambda \bar{C})^{-1} U (\lambda \bar{C})^{-1} W\big) = \lambda^{-2} \langle U, W \rangle_{\bar{C}}$, so with Eq. 19

$$\langle u_i', u_j' \rangle_{\bar{C}'} = \lambda^{-2} \langle \lambda u_i, \lambda u_j \rangle_{\bar{C}} = \langle u_i, u_j \rangle_{\bar{C}}.$$

Norms and angles follow as in Steps 7–8 of Prop. 1. □

Constancy of c_i along the trajectory is what makes the added term of Eq. 17 vanish under differentiation; a gain that varies along the trajectory contributes to the velocity and is not absorbed.

B THE DEPTH PROFILE AT THE STIMULUS LEVEL

Here, we resolve the depth profile of Sec. 4.1 into one curve per image. The published profile correlates the whole angular matrix at each depth, so it reports where the shared geometry is strongest on average rather than for a single image. Correlating each image row across systems gives one value per image and depth,

$$P[i, \ell] \;=\; \mathrm{mean}_t\; \mathrm{corr}\big(\mathrm{row}_i \Theta^B(t),\, \mathrm{row}_i \Theta^L(\ell)\big), \tag{20}$$

computed over the ten EEG subjects. The pooled curve averages images whose own peaks differ widely rather than describing one. Llama's per-image peaks are spread over the whole depth axis: 62 images peak in the first third of depth and 75 in the last, with 63 in between, so the pooled peak at

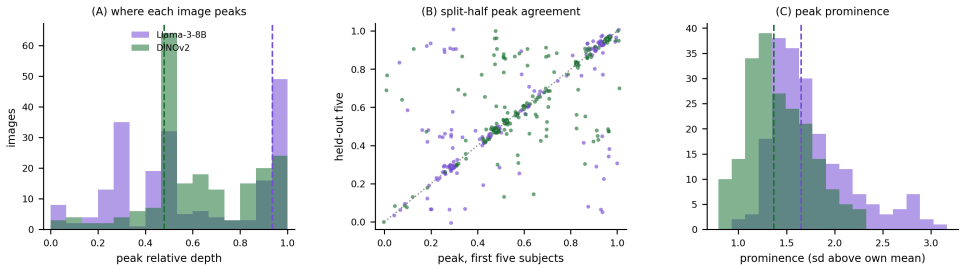


Figure 5: The pooled depth profile resolved per stimulus (ten EEG subjects). **(A)** Where each image's own profile peaks, with the pooled peak dashed. Llama's peaks spread over the whole depth axis, with the largest group at the last blocks; DINOv2's are concentrated at mid-depth, matching its pooled peak. **(B)** Peak depth estimated from five subjects against the held-out five, jittered for legibility. Points are dense on the identity line and scattered off it: most images reproduce their peak exactly, a minority do not. **(C)** Prominence of each image's peak, in standard deviations above its own profile mean, medians dashed. The profile itself is reliable at 0.973 and 0.961.

94% describes only part of the image population (Figure 5A). DINOv2 is concentrated at mid-depth, 110 of 200 images in the middle third, and its pooled peak at 48% sits inside that mode. The profiles themselves are reliable across participants, split-half over subjects giving 0.973 and 0.961. Peak location is the weaker quantity, and it fails in a specific way. Estimated from two disjoint halves of the subject pool, peaks agree at 0.699 and 0.649, with a median displacement of zero and same-third agreement for 81% and 80% of images. Most images reproduce their peak exactly (Figure 5B).

C Category grain: peak locations and order test

Here, we ask two things about the category grains of Sec. 4.4: where along model depth each grain sits, and whether restricting the order test to a single grain recovers the ordering Sec. 4.2 does not find. The two grains peak at different depths, and the within-category peak is set by two categories rather than by the grain. The across-category profile peaks at 48% of network depth in EEG against 65% in MEG for DINOv2, and at 94% against 35% for Llama. The within-category profile appears not to replicate: 52% against 22% for DINOv2, 45% against 32% for Llama, but in each cohort about 77% of the within-category pairs come from that cohort's two largest categories (food, animal), and the two datasets carry different test images under different taxonomies: 20 categories over 200 images in EEG, 64 in MEG, of which 42 are singletons contributing no within-category pair at all. Dropping the two largest categories does not reconcile the cohorts: the within-category peak moves to 30% (EEG) and 87% (MEG) for DINOv2 and to 23% and 0% for Llama, so the location of the within-category peak is not a stable property of either model. Dropping those two categories also leaves the within-category level above the across-category level at every depth in both cohorts for DINOv2, and on average for Llama. Restricting the order test of Sec. 3.6 to a single grain does not recover a significant positive ordering: no DINOv2 split clears the depth-permutation null, and the two Llama splits that do are negative (within-category $\rho = -0.334$, $p = 0.002$; across-category $\rho = -0.201$, $p = 0.045$; Figure 6).

D Reliability of velocity and acceleration

The neighborhood covariance of each system genuinely moves, and in EEG the motion is reliably estimable (Figure 7A). Across the ten EEG subjects, the split-half reliability of the trajectory velocity is 0.528 (bootstrap 95% CI over subjects $[0.453, 0.601]$; per-subject 0.32–0.70) and of acceleration 0.324 ($[0.255, 0.394]$), against 0.357 for position. These are moderately reliable, repeatable descriptors rather than construction noise, although the spread across subjects is wide. The model exhibits the depth analog of motion, but not a monotone deceleration: its layer-to-layer velocity magnitude is unrelated to depth for Llama (Spearman -0.02 on the EEG images, 0.00 on the MEG images) and increases with depth for DINOv2 ($+0.67$ and $+0.64$). So, the EEG trajectories have a reproducible velocity and acceleration.

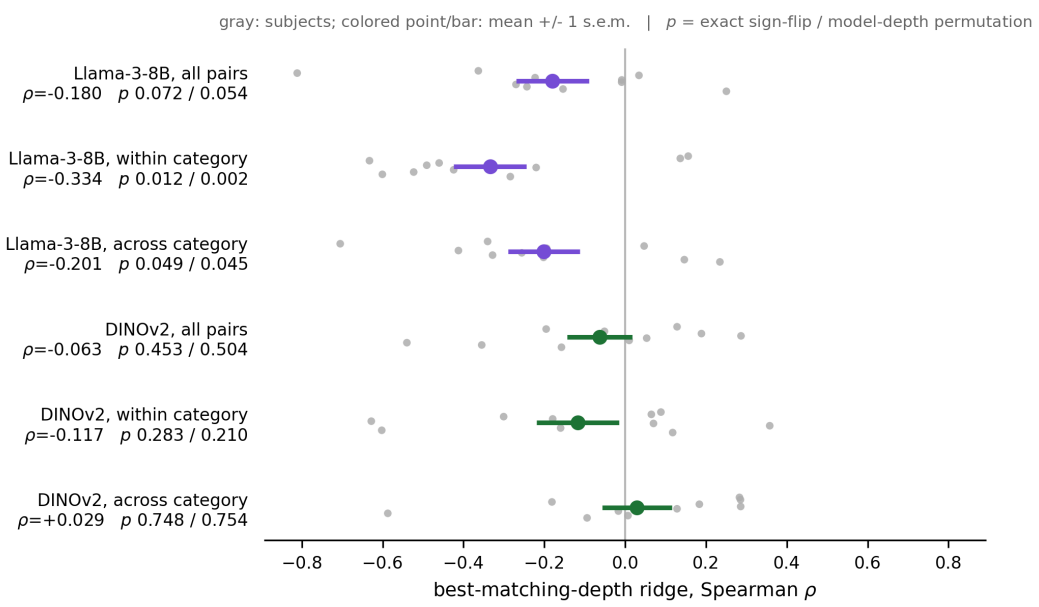


Figure 6: The order test split by pair type, single occipito-temporal SPD, ten EEG subjects. Each row is the ridge statistic of Sec. 3.6 (Spearman correlation between window time and best-matching depth) for one model under one restriction of the pair set: gray points are individual subjects, the colored point and bar the group mean and ± 1 s.e.m. Correlations are recomputed within the restricted set. The two p-values are the exact sign-flip test across subjects and the depth-permutation test of Sec. 3.6, which excludes a model ordering that happens to suit the brain data; no split shows a significant positive ordering, and the Llama within- and across-category ridges are significantly negative.

In MEG, where each half averages six of the twelve repetitions per image (odd against even sessions; forty of eighty in EEG), the descriptors are less reliable and vary strongly across the four subjects: velocity 0.342 (bootstrap 95% CI $[0.048, 0.636]$; per-subject 0.04–0.70), acceleration 0.184 ($[0.021, 0.371]$) and position 0.231 (Figure 7A). The image$\times$image velocity-angle matrix, on which the concordance is built, is comparably reliable in the two cohorts: 0.439 in EEG (bootstrap 95% CI over subjects $[0.358, 0.507]$) and 0.483 in MEG ($[0.361, 0.686]$), where it is more reliable than any individual descriptor.

Panels B and C ask what these dynamics carry about the model, and how much of it the shared neighborhood graph already supplies. Velocity direction carries excess dependence about the model beyond speed in both cohorts (Gaussian-CCA score, Sec. 3.7; per-layer image-permutation $p = 0.005$, the smallest attainable with 200 shuffles, in every layer and subject): 1.78 bits in EEG and 1.94 in MEG, averaged over the eight sampled Llama layers and positive in every subject. Replacing the model with random features of the same dimensionality, pushed through the same neighborhood graph and reduction, lowers this only to 1.56 and 1.69 bits, so most of the dependence does not require the model's own features. Panel C shows the same with the rank concordance of App. J. The true pairing is far above an image-scrambled floor (0.235 against 0.000 in EEG, 0.270 against 0.000 in MEG), but the featureless model concords with cortex more strongly than Llama does, in every subject (0.330 and 0.286). This control is stricter than the graph-scaffold null of App. J and Table 1, which replaces the features of both systems (0.079 and 0.066) and so also removes the graph structure already present in cortex; here only the model is replaced. The true-pairing values differ slightly from App. J (0.225 and 0.271) because panel C uses the eight sampled layers rather than all 32. Measured by per-slice rank concordance, then, the true Llama layers do not improve on a model with no features of its own.

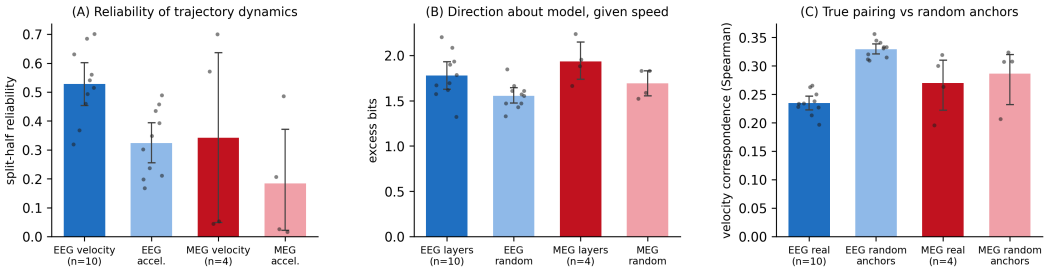


Figure 7: Velocity and acceleration are present and informative. All panels use every subject of both cohorts (ten EEG, four MEG), with bars showing group means, whiskers bootstrap 95% confidence intervals over subjects, and points individual subjects. (A) Split-half reliability of trajectory velocity and acceleration: moderate in EEG (velocity 0.528, acceleration 0.324) and lower and more variable in MEG (velocity 0.342, acceleration 0.184). (B) Gaussian-CCA dependence of velocity direction about the model given speed (excess bits, corrected estimator), for the eight sampled Llama layers against five random-feature controls of the same dimensionality. The real layers give 1.78 bits in EEG and 1.94 in MEG; the controls give 1.56 and 1.69; all values are excess over the estimator's image-permutation null. (C) Velocity correspondence (rank concordance over the eight sampled layers) between cortex and either the true Llama layers or random-feature anchors, a featureless model built on the same neighborhood graph: the random anchors exceed the true pairing in every subject (0.330 against 0.235 in EEG, 0.286 against 0.270 in MEG); an image-scrambled pairing gives 0.000 in both. Unlike the nulls of App. J, only the model is replaced here.

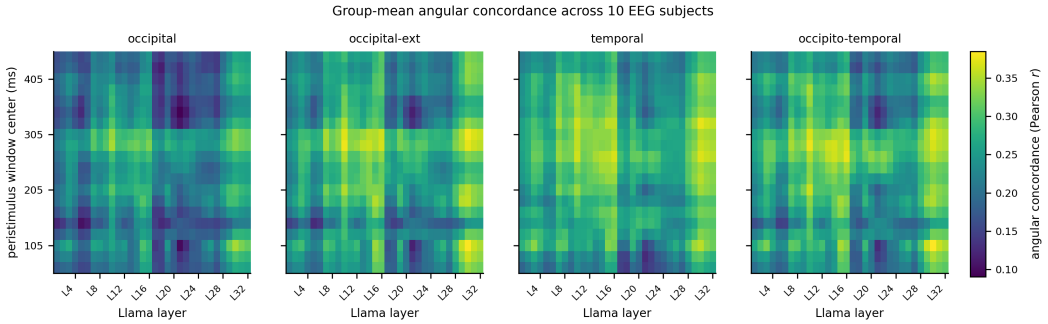


Figure 8: Preserved angular geometry, group mean across ten EEG subjects. Angular concordance (Eq. 7) between cortical velocity directions at each peristimulus window (rows) and Llama layer velocity directions (columns), averaged over subjects, for four channel groups (occipital: SPD(7), only seven channels). Concordance is positive in *every* group-mean cell and each subject's mean is positive (occipito-temporal grand mean 0.2632). Because only within-system relative angles are compared, no common frame or learned map is involved.

E THE ANGULAR GEOMETRY OF THE MOTION IS PRESERVED ACROSS SYSTEMS

Sec. 4.1 reports the concordance as a group mean. This appendix compares the two models on the same footing. Across the ten EEG subjects, the group mean concordance with Llama is 0.2632 (bootstrap 95% CI $[0.2509, 0.2743]$; per-subject 0.2214–0.2917), positive in $10/10$ subjects and positive in *every* cell of the group-mean window×layer grid, which range from 0.1227 to 0.3847 (Figure 8). Both models are examined over their full depth axes, Llama's thirty-two layers and DINOv2's twenty-four, against the same twenty windows. On that footing, the vision model scores higher: DINOv2's average is 0.3913 against Llama's 0.2632, a gap of 0.128.

F COMPARING RECORDING MODALITIES WITH THE FOUNDATION MODELS

In this section, we compare the trace geometry of the EEG/MEG datasets. We restrict the analysis to the 28 THINGS concepts present in both test sets and pair the two cohorts by concept rather than by image, at the cost of a small n. Across all 10×4 EEG–MEG subject pairs the mean concordance is 0.233, positive in $10/10$ EEG subjects and $4/4$ MEG subjects, with every subject pair significant against its own permutation null. Because each EEG subject enters four pairs and each MEG subject ten, the pairs are not independent and we bootstrap over each subject axis separately rather than over pairs: 95% CI $[0.225, 0.240]$ over EEG subjects and $[0.186, 0.264]$ over the four MEG subjects, the latter being the binding constraint. The effect is stable across EEG subjects (0.208–0.251).

Scored on the same 28 concepts, the brain–model agreement is 0.237 (EEG) and 0.301 (MEG) for Llama, and 0.371 and 0.350 for DINOv2, so the EEG–MEG agreement (0.233) is at the level of EEG–Llama and below the other three brain–model edges. We do not read this as evidence that the model is closer to cortex than cortex is to itself. The two modalities come from different participants and are paired by concept, and each cohort's neighborhoods are drawn from its own test set, so the EEG–MEG edge does not share the neighborhood graph that every brain–model edge shares (Sec. 4.3, App. J). It is also attenuated by measurement noise on both of its sides (split-half reliability of the angle matrices 0.439 in EEG and 0.483 in MEG; App. D), whereas a model's representation is deterministic, so a brain–model edge is attenuated on one side only.

The ordering is not a property of the language model. On the same 28 concepts, DINOv2 gives 0.371 in EEG and 0.350 in MEG, above Llama on both instruments and above the brain–brain edge (on all 200 images, 0.3913 and 0.3189). Whatever produces the gap therefore belongs to the comparison rather than to one model class: a label-fed language model and an image-fed vision model—differing in input, architecture, objective and training data—both reach or exceed the agreement between the two recording modalities. Differences in reliability do not explain how the model edges differ between the two instruments: the angle matrices are about as reliable in MEG as in EEG (0.483 against 0.439), yet on all 200 images the MEG edge sits below its EEG counterpart for the vision model (0.3189 against 0.3913) and above it for the language model (0.2835 against 0.2632), and the same holds on the 28 shared concepts (0.350 against 0.371; 0.301 against 0.237).

G EFFECT OF POOLING THE SLICES

The concordance of Sec. E is evaluated slice by slice and order-independent; so it cannot say whether the shared structure survives when each image's covariance trace is treated as a *single object* rather than a bag of slices. Following the rate-invariant framework for covariance trajectories (Zhang et al., 2018), we represent each trace by its transported square-root vector field and compare two traces by geodesic distance in the quotient of that representation by the reparameterization group; this yields, within each system, an image$\times$image matrix of whole-trajectory distances, and the cross-system comparison correlates the two matrices exactly as in Sec. 3.5.

Two implementations are used and reported separately, since they are different statistics. The ***SRVF*** statistic represents each trace by its transported square-root velocity field before taking geodesic distances, and carries the rate-invariance comparison below. The ***flattened-trace*** statistic flattens the tangentized trace and takes Euclidean distances between images directly, and carries the scrambling and slice-matching controls. Computed from one brain input so that the two are directly comparable, they give 0.543 and 0.473 over the ten EEG subjects. Each is compared against its own matched per-slice baseline: 0.247 state-based for the flattened trace, 0.263 velocity-based for the SRVF. The velocity baseline is the primary-pipeline value of Sec. E itself, and the state baseline lies slightly below it. Both pooled forms exceed both baselines in every subject and are significant against an image-permutation null, which sits at 0.000 (Figure 9): the pooled trace concords more strongly than any of its individual slices.

The rate-invariance is not the source of this gain: the fixed-alignment SRVF statistic (0.543) and its rate-invariant (warped) counterpart (0.539) are close (Figure 9A), and in MEG the warped form is lower still (0.540 against 0.560), so quotienting out execution rate does not improve the correspondence in either modality.

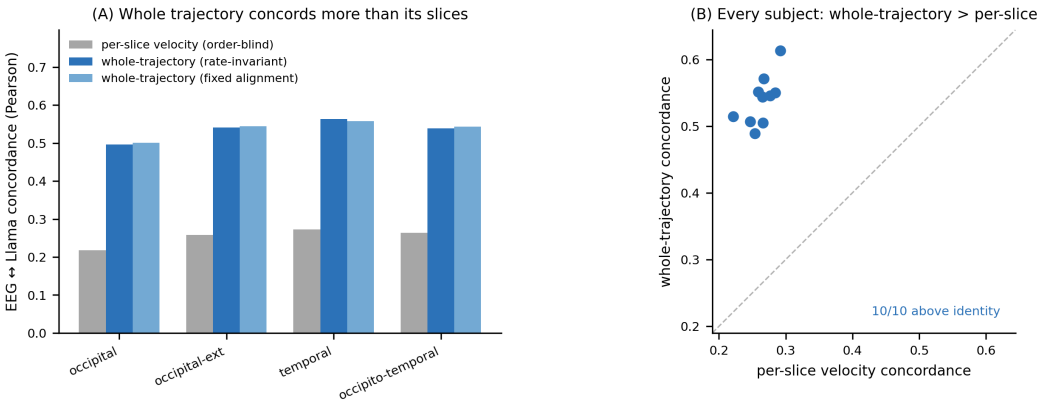


Figure 9: The whole trajectory carries more correspondence than any single slice (EEG, ten subjects). (A) EEG↔Llama concordance in each channel group under three representations of the same trajectories: the per-slice velocity statistic (gray; 0.263 occipito-temporal), the rate-invariant (warped) SRVF whole-trajectory statistic (0.539) and its fixed-alignment counterpart (0.543). (B) Per subject, occipito-temporal: rate-invariant whole-trajectory against per-slice concordance; every subject lies above the identity line.

The comparison replicates in MEG, where a control isolates *why* the pooled trace concords more strongly than its slices. Across the four MEG subjects the pooled concordance is 0.560, positive and significant in all four and above its matched baselines (0.257 state-based, 0.284 velocity-based); the warped form is again lower (0.540). A pooled trace carries position, velocity and acceleration together, so a higher concordance could reflect static geometry accumulated along the trace rather than shared *dynamics*. A shuffled-order control separates them: independently permuting each brain trace's slice order preserves the set of states but destroys their matching across images, and collapses the flattened-trace concordance from 0.343 to 0.159, below the aligned value in every subject (Figure 10). EEG has similar results: 0.473 aligned against 0.081 scrambled, a gain of 0.392 against MEG's 0.184, positive in every subject of both cohorts.

The vision model behaves the same way (0.722 aligned, 0.159 scrambled, same ten subjects). A flattened statistic need not be invariant to the length of the depth axis, and the two models are examined at 24 points against 32, so we tested that directly: sampling Llama at eight depth points instead of all thirty-two moves the flattened trace from 0.473 to 0.495, a difference of 0.022 over a fourfold change in sampling. On that basis, the two models can be compared, and DINOv2 is the higher of the two (0.722 against 0.473 in EEG, 0.496 against 0.343 in MEG), the same ordering as the per-slice statistic of Sec. E, where the vision model is also higher.

The above control does *not* establish a shared temporal order. The per-image scramble destroys two things at once: the sequence within each trace, and the *matching* of slices between traces. Only the second is responsible for the collapse. Applying a *single* permutation to every image, which scrambles the sequence but preserves the matching, leaves the statistic unchanged, in every subject of both cohorts. The statistic compares images through distances between flattened traces, and any transformation applied identically to every image is an isometry of that representation.

H THE RATE-MATCHED ORDER TEST

A natural prior expectation from studies of continuous language is an *order-preserving* correspondence in which deeper model layers align with later brain responses (a monotone time→depth sweep) (Raugel et al., 2025; Caucheteux et al., 2023).Testing the best-matching layer $\ell^*(t)$ for such an ordering, we find none; Eq. 8's rate-free test, reported in Sec. 4.2, does find one against Llama. (Figure 11B,C). The ridge is positive in only $2/10$ EEG subjects, with group mean -0.180 and a bootstrap 95% CI of $[-0.356, -0.023]$, below zero (depth-permutation $p = 0.054$). It is strongly subject-variable, spanning -0.811 to $+0.250$, so a single-subject estimate could have supported either conclusion; and consistency of this kind across subjects is not evidence of ordering when

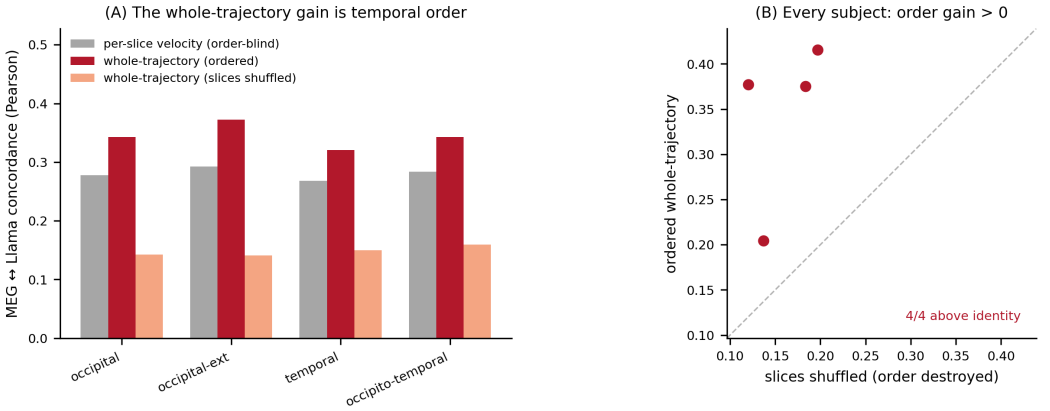


Figure 10: The pooled gain requires slice matching across images (MEG, four subjects). (A) MEG↔Llama concordance in each channel group: per-slice velocity (gray), the aligned pooled trace (red), and the pooled trace with each brain trace's slices independently scrambled (light). Scrambling collapses the concordance below the per-slice level. A single permutation applied to every image, which scrambles sequence but preserves the matching, leaves the statistic unchanged. (B) Per subject: aligned versus scrambled pooled concordance; all four lie above the identity line.

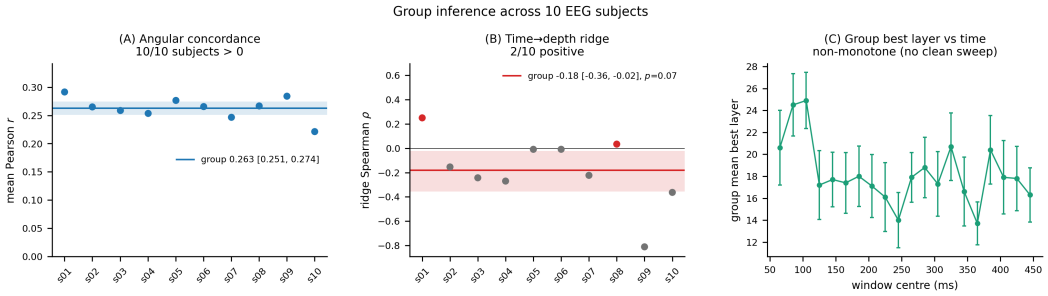


Figure 11: Group inference across ten EEG subjects. (A) Angular concordance: mean correlation per subject, group mean and bootstrap 95% CI; $10/10$ positive. (B) Time→depth ridge: per-subject ridge; $2/10$ positive (red), the rest non-positive (gray), group mean -0.180 $[-0.356, -0.023]$; the p in the panel is the exact sign-flip test, and the depth-permutation p is 0.054. (C) Group-averaged best-matching layer versus window center.

every subject is compared against the same model (Sec. 3.6). The group-averaged best-matching layer is not a clean monotone trajectory: it wanders, with its highest values in the early windows (Figure 11C), so the "best layer" is loosely defined throughout.

The natural objection is that a label-fed model has no visual inference to unfold, so the absence may reflect the model rather than the method. Two things bear on this objection. The vision model, tested on an identical footing, gives the same answer on this statistic. And the label-fed model does show an ordering under the rate-free test of Sec. 4.2, so the ridge's verdict is not a consequence of the model having no depth-wise progression to find. Repeating the test against the vision model on an identical footing gives the same answer: the ridge is weak, group mean -0.063 in EEG ($5/10$ positive, bootstrap 95% CI $[-0.214, +0.079]$) and $+0.213$ in MEG ($3/4$; depth-permutation $p = 0.096$). Sec. 4.2 reports the rate-free test, which is the one place any ordering appears.

I THE SHARED STRUCTURE GOES BEYOND CATEGORY

To ask whether anything model-specific is shared beyond category, we replace Llama with a one-hot category model and score it against the brain trajectory. A one-hot category state is static; it

has no velocity direction of its own and cannot enter the angular statistic. We therefore compare it through the cross-validated Gaussian-CCA dependence estimator (Sec. 3.7), relating each brain velocity trajectory to the one-hot category state through the identical neighborhood-covariance and tangentization pipeline. Measured in this way, the one-hot model reaches a held-out first canonical correlation of 0.92 with the EEG trajectory and 0.93 with MEG, each a group mean over subjects and each significant against the permutation null in every subject of both cohorts ($10/10$ EEG, $4/4$ MEG, 200 permutations). The real brain↔Llama velocity coupling scored the same way is 0.95 in EEG and 0.96 in MEG, also significant in every subject. The one-hot model therefore falls short of Llama in both cohorts: 0.92 against 0.95 in EEG (lower in $10/10$ subjects) and 0.93 against 0.96 in MEG (lower in $4/4$). Scored identically, DINOv2 reaches 0.97 in EEG and 0.95 in MEG, above the one-hot model in every subject of both cohorts. Both models therefore share dependence with the brain trajectory beyond what a category label supplies.

The *angular* statistic is likewise not reducible to a discrete category: partialling category membership out of the EEG angular concordance leaves 0.189 of the 0.263 measured against Llama, and 0.327 of 0.391 against DINOv2. What survives the partial is either genuinely sub-category geometry or the shared CLIP-neighborhood graph on which both pipelines are built.

The neighborhood covariances of both systems are built over the same CLIP neighborhoods, so a shared graph could impose concordance before any neural or model geometry enters. A random-feature null quantifies this directly: replacing both systems' features with independent Gaussian noise while retaining the real graph yields a concordance of 0.079 in EEG and 0.066 in MEG, against observed values of 0.225 and 0.271 on the same data, a margin of $+0.146$ and $+0.205$ against Llama, positive in every subject of both cohorts. Reshuffling the graph independently for the two systems drops the null to 0.000 and 0.001, confirming that the shared neighborhood structure, not the covariance construction as such, is what the null captures.

J THE GRAPH-SCAFFOLD NULL

Both brain and model neighborhood covariances are built over the same $k = 20$ CLIP neighborhoods, so the shared graph can impose concordance before any neural or model geometry is measured. This appendix gives the mechanism and the measured null.

We replace each system's features with independent random vectors x_J but keep the real CLIP neighborhoods. The neighborhood covariance $C(I) = \frac{1}{k}\sum_{J\in N_k(I)}(x_J - \bar{x}_I)(x_J - \bar{x}_I)^\top$ equals, in expectation over the random draw, the population covariance for every I; the *image-specific* part is the sampling fluctuation fixed by which vectors lie in $N_k(I)$. Two images I, I' share the vectors of $S = N_k(I) \cap N_k(I')$, so their whitened tangent coordinates share that component and split as

$$t_A(I) = g\big(N_k(I)\big) + \varepsilon_A(I), \qquad t_B(I) = g\big(N_k(I)\big) + \varepsilon_B(I),$$

where g is the overlap-driven component common to two independently randomized systems and $\varepsilon_A, \varepsilon_B$ are independent. Both systems' angle matrices are then noisy read-outs of the *same* overlap matrix $O(I, I') = |N_k(I) \cap N_k(I')|/k$, and the graph-null concordance is the reliability of that structure,

$$\rho_{\text{graph-null}} \;\approx\; \frac{\mathrm{Var}[g(O)]}{\mathrm{Var}[g(O)] + \mathrm{Var}[\varepsilon]}.$$

It is non-zero whenever O has block structure, and in this graph that block structure is largely category co-membership: same-category image pairs share 8.2 of their 20 neighbors on average against 2.2 for different-category pairs, and 59% of all pairs have non-empty overlap.

Table 2 gives the outcome of simulating this directly on all 200 images; each isotropic value is a mean over 20 random-feature draws, and each spectrum-matched value over 5 draws per subject. So that observed and null values are directly comparable, both go through the same path (features to neighborhood covariance to velocity direction to concordance) and are scored with the same rank (Spearman) concordance, which is why the observed value quoted here (0.225 in EEG) differs from the Pearson value of Sec. 4.1 (0.2632); margins are taken within this statistic.

The two null rows are properties of the graph and the reduction, not of the model: neither uses the real model features, and the small drift between model columns follows the number of depth points

Table 2: Graph-scaffold nulls against the observed velocity concordance, all scored with the rank (Spearman) version of Eq. 7 on each cohort's 200 test images. The null rows replace both systems' features with random vectors on the real CLIP neighborhood graph (isotropic: mean of 20 draws; matched to each system's covariance spectrum: 5 draws per subject); the floor also relabels each system's graph independently. Observed values are means over subjects.

	Llama-3-8B		DINOv2 ViT-L/14	
	EEG (10)	MEG (4)	EEG (10)	MEG (4)
Independent graphs (floor)	0.000	0.001	0.000	0.000
Shared graph, isotropic features	0.079	0.066	0.081	0.069
Shared graph, spectrum-matched features	0.126	0.111	0.220	0.207
Observed	0.225	0.271	0.369	0.291

(32 against 24) rather than the architecture. Only the observed row, and the spectrum-matched null through the real covariance spectrum, are model-specific.

Reshuffling the graph independently for the two systems collapses the null to the floor, so what the null captures is the shared neighborhood structure rather than the covariance construction as such. Against Llama, the isotropic shared graph alone reproduces 35% of the observed concordance in EEG and 24% in MEG, and the observed value exceeds it by $+0.146$ and $+0.205$, positive in every subject of both cohorts; across the 20 random-feature draws the null itself has standard deviation 0.008 in EEG and 0.009 in MEG, so seed variability is small relative to these margins. Against DINOv2 the margin is larger in both cohorts ($+0.288$ in EEG, positive in $10/10$ subjects; $+0.222$ in MEG, positive in $4/4$).

K Separation order of categories

This section examines the order in which categories separate along a trajectory. Within each system we form a category×category dissimilarity matrix from the angles between category-mean velocity directions, cluster it by average linkage, and annotate every internal node of the resulting tree with an *onset*: the first window at which the between-group minus within-group angle exceeds a permutation threshold on two consecutive windows.

We find that the order is coarse-to-fine (Figure 4). Because the two datasets carry different taxonomies, EEG's 20 broad categories against MEG's 64, each cohort is analysed on its own vocabulary, requiring at least three images per category: EEG (18 categories, 197 images) and MEG (18, 150). Pooling resolved splits over subjects, coarse distinctions (separating six or more categories) reach threshold in the earliest window in both cohorts while finer ones appear later: $65 \rightarrow 85$ ms in EEG and $65 \rightarrow 125$ ms in MEG. The instruments agree on the ordering and differ by two 20 ms windows in where the fine splits land. This progression is consistent with MEG evidence that broad category information emerges before more object-specific semantic information during the first few hundred milliseconds of object processing (Clarke et al., 2013; 2015).

Splits drawn from a nested tree and pooled over subjects are not independent, so we test at the participant level instead: per subject, the median fine-split onset minus the median coarse-split onset, sign-flipped across subjects. The median difference is $+30$ ms (mean $+50$), positive in $6/10$ subjects with the remaining four exactly tied, $p = 0.031$ (exact, all 2^{10} sign patterns), on EEG's own vocabulary. MEG separates in the same direction in three of the four subjects, the fourth tied, median $+65$ ms, but four subjects cannot establish it: the sign-flip null has sixteen points and returns $p = 0.250$. One confound is real but does not appear to explain the effect: an onset is a detection time, and coarse splits rest on more images than fine ones (median 109 against 33 images), with onset correlating with image count at $\rho = -0.33$; but in the only image-count quartile that contains both coarse and fine splits the ordering is preserved ($+60$ ms), and within the inter-quartile band of

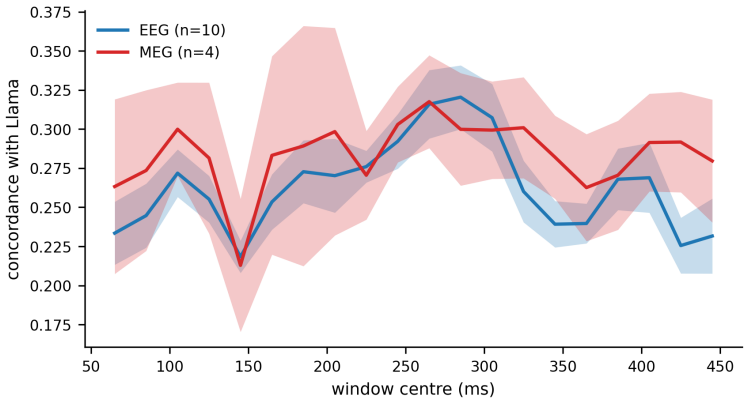


Figure 12: Cross-system concordance is present from the first window in both instruments. Window-by-window velocity-direction concordance with Llama (mean over layers), pooled over the ten EEG subjects and four MEG subjects; shaded bands are ± 1 SEM across subjects. Both curves begin well above the image-permutation null at the first window (~ 65 ms). MEG (red) reaches 90% of its maximum by ~ 105 ms and peaks near 265 ms; EEG (blue) rises to a broad peak near 285 ms, reaching 90% of its maximum by ~ 245 ms.

image counts the medians remain ordered (65 and 75 ms). In MEG the two quartiles containing both kinds of split disagree ($+31$ and -18 ms). We report the progression as an ordinal tendency.

The brain and Llama lay out this structure in broadly similar proportions along their trajectories. In the brain, 65% of resolved splits (57% in MEG) clear threshold in the first window and the remainder resolve over the following ~ 360 ms; in Llama, 60% of resolved splits clear threshold at the first sampled layer step (44% on the MEG images) and the remainder resolve over depth, so the category structure is not fully expressed at the earliest depth we sample, even though Llama receives each image's object label.

We repeat the analysis with the image-fed vision model, which also separates coarse categories before fine ones. On the EEG images, 42% of its resolved splits are already significant at the first depth step (Llama: 60%), and the rest resolve deeper, the last at 61% of depth (Llama: 52%). In both models, coarse splits resolve at a median depth step of 0 (the first block) and fine splits at a median of 3. On the MEG images, where both models are scored on the same 18 categories, both resolve 44% of splits at the first step and separate coarse before fine (median steps 0 and 5 for DINOv2, 0 and 6.5 for Llama). Both models therefore show a descriptive coarse-to-fine ordering over depth, whichever input format they receive.

The question of which foundation model builds the category tree the brain builds is a separate question, and the two datasets agree on it. With both models reduced identically, the vision model's tree is *further* from the brain's than the language model's on both image sets (cophenetic agreement 0.928 against 0.949 on the EEG images, 0.851 against 0.895 on the MEG images). The brain's tree *topologies* are thus not better matched by the image-fed model.

The same timeline can be read from the cross-system agreement rather than from the tree: the window-by-window concordance with Llama already stands above its image-permutation null in the first analysis window in both instruments, and EEG rises to 90% of its maximum by ~ 245 ms (Figure 12).

L Comparison against standard methods

We compare the angular concordance with four standard alternatives: orthogonal Procrustes, linear centered kernel alignment (CKA), representational similarity analysis (RSA), and dynamical similarity analysis (DSA). The rate-invariant whole-trajectory (SRVF) statistic is treated in App. G. Each comparison below turns on what the method can be asked to compute, not on how well it scores. Orthogonal Procrustes has no pairwise object at all. It truncates the two systems to a common

dimensionality and then fits a single orthogonal map over all N images jointly, so there is no image-pair term to restrict and no per-image term free of that global fit. Three of the analyses reported here are therefore unavailable to it in any form: the depth profile split into within- and across-category pairs (Figure 3), the order test restricted to a single grain (Figure 6), and the depth profile resolved per stimulus (Figure 5), which carries image i by its row of Θ. It is excluded on a second ground from any claim whose content is that no cross-system map is fitted, since fitting that map is what it does.

Linear centered kernel alignment does have a pairwise form: its numerator $\|A^\top B\|_F^2$ is a sum over image pairs of two centered Gram matrices, so the same restrictions can be written down. They do not mean the same thing. Every Gram entry is an inner product taken about the mean over *all* N images, so a CKA restricted to the within-category pairs still carries the across-category images in its centering, and is not the statistic computed within that subset; the same holds of a single row. The angular statistic re-centers inside whatever set of pairs it is given, so Figures 3, 6 and 5 report the comparison within each grain rather than a masked view of the whole. CKA is in addition invariant to isotropic scaling of a system as a whole but not to rescaling each image separately.

Representational similarity analysis is in the same class as the angular statistic: both form an image$\times$image matrix and compare its entries, so both support the analyses above; with a Pearson correlation, Eq. 7 is exactly an RSA comparison of two such matrices. The difference lies in what fills the matrix and in what is read off the grid. An RSA matrix holds dissimilarities between the response patterns to two images, the geometry of states; ours holds angles between the velocity directions of their neighborhood-covariance trajectories, the geometry of motion. RSA answers the geometry question slice by slice and stops there; the order statistic of Eq. 8 is defined on the same window$\times$depth grid and asks a question RSA is not used to ask, whether the two systems traverse that geometry in a corresponding order.

Dynamical similarity analysis is the only one of these methods that asks the order question, and it asks it differently. It fits a linear (Koopman) operator to each system's trajectory and compares the two operators, which requires bringing them into a common basis; Eq. 8 fits nothing and compares no operators. The assumptions differ accordingly: a Koopman fit needs an operator that is stable along the axis and estimable from the transitions available, here nineteen on the brain axis and thirty-one or twenty-three on the model side, where the warped path needs only that both axes be traversed monotonically and assumes nothing about rate. We report no comparison against it because the implementation available to us compares the two operators' eigenvalue spectra rather than aligning them, and that alignment is the step at issue.